VLDB JOURNAL manuscript No.
(will be inserted by the editor)

A Look Ahead Approach to Secure Multi-party Protocols

Mehmet Ercan Nergiz · Ercüment Çiçek · Yücel Saygın

Received: date / Accepted: date

Abstract Secure multi-party protocols have been proposed to enable non-colluding parties to cooperate without a trusted server. Even though such protocols prevent information disclosure other than the objective function, they are quite costly in computation and communication. Therefore, the high overhead makes it necessary for parties to estimate the utility that can be achieved as a result of the protocol beforehand. In this paper, we propose a *look ahead* approach, specifically for secure multi-party protocols to achieve distributed k-anonymity, which helps parties to decide if the utility benefit from the protocol is within an acceptable range before initiating the protocol. Look ahead operation is highly localized and its accuracy depends on the amount of information the parties are willing to share. Experimental results show the effectiveness of the proposed methods.

Keywords Secure multi party computation · Distributed k-anonymity · Privacy · Security

1 Introduction

Secure multi party computation (SMC) protocols are one of the first techniques for privacy preserving data mining in distributed environment [19]. The idea behind these protocols is based on the theoretical proof that two or more parties, both having their own private data, can collaborate to calculate any function on the union of their data [7]. While doing so, the protocol does not reveal anything other than the output of the function and does not require a trusted third party. While this property is promising for privacy preserving applications, SMC may be prohibitively expensive. In fact, many SMC protocols for privacy preserving data mining suffer from high computation and communication costs. Furthermore, those that are closest to be practical are based on *semi-honest model*, which assumes that parties will not deviate from the protocol. Theoretically, it is possible to convert semi-honest models into *malicious models*. However, resulting protocols are even more costly.

The high overhead of SMC protocols raises the question whether the information gain (increase in utility) after the protocol is worth the cost. This is a valid argument for mining on horizontally or vertically partitioned data (but especially crucial for horizontally partitioned data where objective function is well defined on the partitions since they have the same schema.). More specifically, for private table T_σ of party P_σ and an objective function O; initiating the SMC protocol is meaningful only if the information gain from O; $|I_\sigma| = |I(O(T_\cup)) - I(O(T_\sigma))|$ where $T_\cup$ is the union of all private tables, is more than a user defined threshold c. Of course $|I_\sigma|$ cannot be calculated without executing the protocol. However it may be possible to estimate it by knowing some prior (and non-sensitive) information about $T_\cup$.

To the best of our knowledge, this is the first work that *looks ahead* of an SMC protocol and gives an estimate for I_σ. We state that an ideal look ahead satisfies the following:

1. Methodology is highly localized in computation, it is fast and requires little communication cost (at least asymptotically better than the SMC protocol).
2. Methodology relies on non-sensitive data, or better, data that would be implied from the output of the objective function.

We state that an ideal look ahead will benefit the parties in answering the following:

1. How likely the information gain I_σ will be within an acceptable range?
2. Since efficiency of SMC depends heavily on data, what size of private data would be enough to get an acceptable I_σ?

Our focus is the SMC protocol for distributed k-anonymity previously studied in [31,11,10]. k-Anonymity is a well known privacy preservation technique proposed in [27,24] to prevent linking attacks on shared databases. A database is said to be k-anonymous if every tuple appears in the database at least k times. k-Anonymization is the process of enforcing k-anonymity property on a given database

M. E. Nergiz
Sabanci University, Istanbul, Turkey
Tel.: +90 216 483 9000 - 2114
E-mail: ercann@sabanciuniv.edu

E. Çiçek
Sabanci University, Istanbul, Turkey
E-mail: ercumentc@su.sabanciuniv.edu

Y. Saygın
Sabanci University, Istanbul, Turkey
Tel.: +90 216 483 9576
E-mail: ysaygin@sabanciuniv.edu

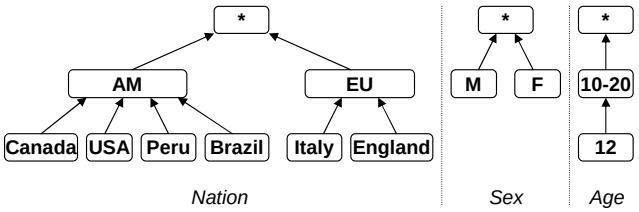


Fig. 1 DGH structures

by using generalization and suppression of values. Works in [11,10] assume that data is vertically partitioned among two parties and they share a common key making a join possible. Authors in [11] propose a semi-honest SMC solution to create a k-anonymization of the join without revealing anything else (The protocol takes around 2 weeks time to execute for $k = 100$ and 30162 tuples.). Work in [31] assumes horizontally partitioned data.

The motivation behind k-anonymity or distributed k-anonymity as a privacy notion has been studied extensively in the literature. Many extensions to k-anonymity has been proposed that address various weaknesses of the notion against different types of adversaries [8,18,20,22,29,30,21,3]. ℓ-Diversity [20] is one such extension that enforces constraints on the distribution of the sensitive values. We first focus on the k-anonymization process and show later how the proposed methodology can be extended for ℓ-diversity. Our contribution can be summarized as follows:

1. We design a fast look ahead of distributed k-anonymization that bounds the probability that k-anonymity will be achieved at a certain utility. Utility is quantified by commonly used metrics from the anonymization literature.
2. Look ahead works for horizontally, vertically and arbitrarily partitioned data.
3. Look ahead exploits prior information such as total data size, attribute distributions, or attribute correlations, all of which require simple SMC operations. Look ahead returns tighter bounds as the security constraints allow more prior information.
4. We show how look ahead can be extended to enforce diversity on sensitive attributes as in [18,20].
5. To the best of our knowledge, this work is the first attempt in making a probabilistic analysis of k-anonymity given only statistics on the private data.

2 Background

2.1 k-Anonymity and Table Generalizations

Given a dataset (table) T, $T[c][r]$ refers to the value of column c, row r of T. $T[c]$ refers to the projection of column c on T and $T[.][r]$ refers to selection of row r on T. We write $|t \in T|$ for the cardinality of tuple $t \in T$.

Although there are many ways to generalize a given data value, in this paper, we stick to generalizations according to domain generalization hierarchies (DGH) given in Figure 1 since they are widely used in the literature.

Definition 1 (i-Gen Function) For two data values v^* and v from some attribute A, we write $v^* = \Delta_i(v)$ if and only if v^* is the ith parent of v in the DGH for A. Similarly for tuples t, t^*, $t^* = \Delta_{i_1,\cdots,i_n}(t)$ iff $t^*[c] = \Delta_{i_c} t[c]$ for all columns c. Function Δ returns all possible generalizations of a value v. We also abuse notation and write $\Delta^{-1}(v^*)$ to indicate the leaf nodes of the subtree with root v^*.

E.g., given DGH structures in Figure 1. Δ_1(USA)=AM, Δ_2(Canada) =*, $\Delta_{0,1}$(<M,USA >)=<M,AM >, Δ(USA)={USA , AM,*}, Δ^{-1}(AM)= {USA , Canada , Peru , Brazil }

Definition 2 (Single Dimensional Generalization) We say a table T^* is a $\mu = [i_1, \cdots, i_n]$ single dimensional generalization of table T with respect to set of attributes QI={$A_1, \cdots, A_n$} if and only if $|T| = |T^*|$ and records in T, T^* can be ordered in such a way that $T^*[QI][r] = \Delta_{i_1,\cdots,i_n}(T[QI][r])$ for every row r. We say μ is a generalization mapping for T and T^*; and write $T^* = \Delta_\mu(T)$.

Definition 3 (μ-Cost) Given a generalization T^*, μ-cost returns the generalization mapping of T^*: $\mu(T^*) = [i_1, \cdots, i_n]$ iff $T^* = \Delta_{i_1,\cdots,i_n}(T)$

For example, Tables T^*_σ, T^*_1 are [0,2] generalizations of T_σ and T_1 respectively w.r.t. attributes sex and nation. Similarly $T^*_{\cup,\sigma} = \Delta_{0,1}(T_1)$, $T^*_{\cup,1} = \Delta_{0,1}(T_2)$. μ-Cost of $T^*_{\cup,1}$ is $[0,1]$.

Definition 4 Given two generalization mappings μ^1=[$i^1_1, \cdots, i^1_n$] and $\mu^2 = [i^2_1, \cdots, i^2_n]$, we say μ^1 is a higher mapping than μ^2 and write $\mu^1 \subseteq \mu^2$ iff $\mu^1 \neq \mu^2$ and $i^1_j \geq i^2_j$ for all $j \in [1-n]$. We define $\mu^1 - \mu^2 = \sum_j i^1_j - i^2_j$

E.g., [0,2] is a higher mapping than [0,1].

Corollary 1 *Given mappings $\mu^1 \subset \mu^2$ and $T^*_1 = \Delta_{\mu^1}(T)$, $T^*_2 = \Delta_{\mu^2}(T)$; T^*_2 is better utilized (contains more information) than T^*_1.*

The above corollary is true because T^*_1 can be constructed from T^*_2. E.g., $T^*_{\cup,\sigma}$ is better utilized than T^*_σ.

In this paper, without loss of generality, we use single dimensional generalizations. However, underlying ideas can also be applied to multi dimensional generalizations [16]. We now revisit briefly k-anonymity definitions.

While publishing person specific sensitive data, simply removing uniquely identifying information (SSN, name) from data is not sufficient to prevent identification because partially identifying information, quasi-identifiers, (age, sex, nation . . .) can still be mapped to individuals (and possibly to their sensitive information such as salary) by using

Table 1 Home party and remote party datasets and their local and global anonymizations

Name	Sex	Nation	Salary
q1	F	England	>40K
q2	M	Canada	≤40K
q3	M	USA	≤40K
q4	F	Peru	≤40K

T_σ

Name	Sex	Nation	Salary
q1	F	*	>40K
q2	M	*	≤40K
q3	M	*	≤40K
q4	F	*	≤40K

T^*_σ

Name	Sex	Nation	Salary
q5	M	Canada	>40K
q6	M	USA	>40K
q7	F	Brazil	≤40K
q8	F	Italy	>40K

T_1

Name	Sex	Nation	Salary
q5	M	*	>40K
q6	M	*	>40K
q7	F	*	≤40K
q8	F	*	>40K

T^*_1

Name	Sex	Nation	Salary
q1	F	EU	>40K
q2	M	AM	≤40K
q3	M	AM	≤40K
q4	F	AM	≤40K
q5	M	AM	>40K
q6	M	AM	>40K
q7	F	AM	≤40K
q8	F	EU	>40K

$T^*_\cup = T^*_{\cup,\sigma} \cup T^*_{\cup,1}$

external knowledge [26]. (Even though T_σ of Table 1 does not contain info about names, releasing T_σ is not safe when external information about QI attributes is present. If an adversary knows some person Alice is a British female; she can map Alice to tuple q1 thus to salary `>40K` .) The goal of privacy protection based on k-anonymity is to limit the linking of a record from a set of released records to a specific individual even when adversaries can link individuals via QI:

Definition 5 (k-Anonymity [26]) A table T^* is k-anonymous w.r.t. a set of quasi identifier attributes QI if each record in $T^*[QI]$ appears at least k times.

For example, T^*_σ, T^*_1 are 2-anonymous generalizations of T_σ and T_1 respectively. Note that given T^*_σ, the same adversary can at best link Alice to tuples q1 and q4.

Definition 6 (Equivalence Class) The equivalence class of tuple t in dataset T^* is the set of all tuples in T^* with identical quasi-identifier values to t.

For example, in dataset T^*_1, the equivalence class for tuple $q1$ is $\{q1, q4\}$.

There may be more than one k-anonymizations of a given dataset, and the one with the most information content is desirable. Previous literature has presented many metrics to measure the utility of a given anonymization [9, 23, 13, 4, 1]. We revisit Loss Metric (LM) defined in [9]. LM penalizes each generalization value v^* proportional to $|\Delta(v^*)|$ and returns an average penalty for the generalization. Let a is the number of attributes, then:

$$LM(T^*) = \frac{1}{|T| \cdot a} \sum_{i,j} \frac{|\Delta(T[i][j])| - 1}{|\Delta(*)| - 1}$$

Since k-anonymity does not enforce constraints on the sensitive attributes, sensitive information disclosure is still possible in a k-anonymization. (e.g., in T^*_1, both tuples of equivalence class {q2,q3} have the same sensitive value.) This problem has been addressed in [20, 18, 8] by enforcing diversity on sensitive attributes within a given equivalence class. We show in Section 6 how to extend the look ahead process to support diversity on sensitive attributes. For the sake of simplicity, from now on we assume datasets contain only QI attributes unless noted otherwise.

2.2 Distributed k-Anonymity

Even though k-anonymization of datasets by a single data owner has been studied extensively; in real world, databases may not reside in one source. Data might be horizontally or vertically partitioned over multiple parties all of which may be willing to participate to generate a k-anonymization of the union. The main purpose of the participation is using a larger dataset to create a better utilized k-anonymization.

Suppose in Table 1, two parties P_σ and P_1 have T_σ and T_1 as private datasets and agree to release a 2-anonymous union. Since data is horizontally partitioned, one solution is to 2-anonymize locally and take a union. T^*_σ, T^*_1 are optimal (with minimal distortion) 2-anonymous full-domain generalizations of T_σ and T^*_1 respectively. However, optimal 2-anonymization of $T_\sigma \cup T_1$; $T^*_\cup$ is better utilized than $T^*_\sigma \cup T^*_1$. So there is a clear benefit in working on the union of the datasets instead of working separately on each private dataset.

As mentioned above, in most cases, there is no trusted party to make a secure local anonymization on the union. So SMC protocols are developed in [11, 10, 31] among parties to securely compute the anonymization with semi-honest assumption.

We assume data is horizontally partitioned but we will state how to modify the methodology to work on vertically partitioned data. We assume we have $n+1$ parties P_σ,P_1, $\cdots$,P_n with private tables $T_\sigma, T_1, \cdots, T_n$. The *home party* P_σ is looking ahead of the SMC protocol and *remote parties* $P_1, \cdots, P_n$ are supplying statistical information on the union of their private tables, $\bigcup_i T_i$. We use the notation $T_\cup$ for the global union (e.g., $T_\cup = T_\sigma \cup \bigcup_i T_i$). We use the superscript * in table notations to indicate anonymizations. We use the

notation $T^*_{\cup,i}$ to indicate the portion of $T^*_{\cup}$ that is generalized from T_i (see Table 1), thus $T^*_{\cup} = T^*_{\cup,\sigma} \cup \bigcup_i T^*_{\cup,i}$. Until Section 5.7, without loss of generality, we assume $n = 1$.

2.3 k-Anonymity Extensions

Many extensions to k-anonymity have been proposed to deal with potential disclosure problems in the basic definition [8,18,20,22,29,30,21,3]. Problems arise mostly because k-anonymity does not enforce diversity on the sensitive values within an equivalence class. Even though, there is no distributed protocol proposed for the k-anonymity extensions yet, there is strong motivation in doing so. In Section 6, we design a look ahead for recursive (c,ℓ)-diversity protocol.

Definition 7 (Recursive (c,ℓ)-diversity [20]) Let the ordered set $R_i = \{r_1, \cdots r_m\}$ hold the frequencies of sensitive values that appear in an equivalence class EC_i. We say a table T^* is recursive (c,ℓ)-diverse iff for all $EC_i \in T^*$, $r_1 \leq (r_\ell + r_{\ell+1} + \cdots + r_m)$.

From now on, without loss of generality, we assume we have only two values in the sensitive attribute domain ($m = 2, \ell = 2$). In Table 1, $T^*_{\cup}$ is $(0.5,2)$-diverse since for all equivalence classes, the frequencies of $\leq$`40K` and `>40K` are the same (i.e., $r_1 = r_2$). However T^*_{σ} does not respect any diversity requirement (except when $c = 0$), since all tuples in equivalence class $\{q_2, q_3\}$, have salary $\leq$`40K`.

3 Information Gain

Given the cost of most SMC protocols, there arises the need to justify the information gain from the protocols. Surely, such gain is nonnegative, but could be 0 or may not meet the expectations. So it is imperative for collaborating parties to decide if information gain is within acceptable range:

Definition 8 (Info Gain) Let $P_\sigma, P_1, \cdots, P_n$ be $n+1$ parties with private tables $T_\sigma, T_1, \cdots, T_n$. Let O be the objective function for the SMC protocol and I be the utility function (information content) defined on the output domain of O. Local info gain for a single party P_σ is defined as $|I_\sigma| = I(O(T_\cup)) - I(O(T_\sigma))$ where $T_\cup = T_\sigma \cup \bigcup_i T_i$. Global info gain for the protocol is $|I| = \sum_j |I_j| + |I_\sigma|$.

Each party involving in an SMC expects to gain from SMC either locally or globally depending on the application. In this work, we assume that parties require the local info gain to exceed some threshold c before they proceed with the SMC protocol. However, without total knowledge of all private tables ($T_\cup$), parties can only have some *confidence* that SMC will meet their expectations:

Definition 9 (c,p-sufficient SMC) For a party P_σ, an SMC is c,p-sufficient with respect to some prior knowledge K on $\cup_i T_i$, if $\mathcal{P}(|I_\sigma| \geq c \mid K) \geq p$. We say SMC is c,p-sufficient iff it is c,p-sufficient for all parties involved.

Our goal in a look ahead process will be to check if a given SMC is c,p-sufficient for a user defined c and p.

For distributed k-anonymity, the objective function O is trivially the optimal k-anonymization which we name as O_k. Specifically, in this paper, we will make use of single dimensional generalizations to achieve k-anonymity. This generalization technique has been used in many previous work on anonymization [15,20,18,22]. As mentioned above, our work can be extended for multidimensional generalizations [16,22] as well.

Information gain (I) is proportional to the quality of the anonymization. It is challenging to come up with a standard metric to measure the quality of an anonymization [23]. In this work, we will be using the μ-cost as the quality metric. Recall that a higher mapping is less utilized than a lower mapping, and '-' operation has been defined over mappings in Definition 4. μ-cost can be used for horizontally partitioned data.

Calculation of LM cost is possible if we know attribute distributions (denoted with K_F) and the generalization mapping. So there is a direct translation between the μ-cost and LM cost for single dimensional generalizations given K_F. The advantage of translating μ-cost to LM cost is that LM cost can be used for arbitrarily partitioned data. For vertical partitioning, each party has at least one missing attribute. We assume a total suppression ($*$) for data entries from the missing attributes when calculating LM cost.

We can now specialize c,p-sufficiency for distributed k-anonymity problem:

Definition 10 (c,p-sufficient k-Anonymity) For a party P_σ, a distributed k-Anonymity protocol is c,p-sufficient with respect to some prior knowledge K on $\cup_i T_i$, iff

$$\mathcal{P}(\mu(O_k(T_\cup)) - \mu(O_k(T_\sigma)) \geq c \mid K) \geq p$$

We say SMC is c,p-sufficient iff it is c,p-sufficient for all parties involved.

Informally, SMC is sufficient for an involving party if the difference between the optimal generalization mapping for the union and the optimal mapping for the local table is more than c with p probability. Of course, the party can only calculate such a probability if she has some knowledge on the union denoted by K. The amount of prior knowledge K is crucial in successfully predicting the outcome of an SMC. As mentioned before, prior knowledge K cannot be sensitive information. Non-sensitive K can be derived in three ways:

1. Information that could also be learned from the anonymization such as the global dataset size.

2. Statistics about global data that are not considered as sensitive. In the case of k-anonymity, statistics that are not individually identifying such as attribute distributions are acceptable.
3. Based on the assumption that global joint distribution is similar with local distribution, information that can be gained from the local dataset. This type of prior knowledge is the most tricky one since over fitting to local distribution needs to be avoided. Such an information can be in terms of highly supported association rules in the local dataset.

We show, in later sections, how to check for sufficiency of distributed k-anonymity protocol given global attribute distributions which we denote with K_F.

Definition 11 (Global attribute distribution K_F) A distribution function f_c^T for an attribute c is defined over a dataset T such that given a value v^* returns the number of entities t in T with $v^* \in \Delta(t[c])$. Global attribute distribution K_F sent to a home party P_σ contains all distribution function on $\bigcup_i T_i$.

In Table 1, $f^{T_1}_{\texttt{Nation}}(\texttt{AM}) = 3$, $f^{T_1}_{\texttt{Nation}}(\texttt{EU}) = 1$. For the parties $\{P_\sigma, P_1\}$, $K_F = \{f^{T_1}_{\texttt{Sex}}, f^{T_1}_{\texttt{Nation}}\}$.

4 Problem Definition

Given Section 3, distributed k-anonymity protocol is c,p-sufficient for P_σ iff

$$\mathcal{P}(\mu(O_k(T_\cup)) - \mu(O_k(T_\sigma)) \geq c \mid K_F) \geq p$$

$\mu^\sigma = \mu(O_k(T_\sigma))$ requires local input and can be computed by P_σ.

$$\mathcal{P}(\mu(O_k(T_\cup)) - \mu^\sigma \geq c \mid K_F) \geq p$$

Let $S_\mu = \{\mu_1^{=c}, \cdots, \mu_m^{=c}\}$ be the mappings that are exactly c distance beyond μ^σ and $\{\mu_1^{>c}, \cdots, \mu_m^{>c}\}$ be the mappings that are more than c distance beyond μ^σ. Let also A_μ be the event that $\Delta_\mu(T_\cup)$ is k-anonymous. Then we have;

$$\begin{aligned}
&\mathcal{P}(\mu(O_k(T_\cup)) - \mu^\sigma) \geq c \mid K_F) \\
&= \mathcal{P}((\cup_i A_{\mu_i^{=c}}) \cup (\cup_i A_{\mu_i^{>c}}) \mid K_F) \\
&= \mathcal{P}(\cup_i A_{\mu_i^{=c}} \mid K_F) \\
&\geq Max_i\ \mathcal{P}(A_{\mu_i^{=c}} \mid K_F)
\end{aligned}$$

This follows from the monotonicity of k-anonymity. So the problem of sufficiency reduces to prove that, for at least one $\mu \in S_\mu$;

$$\mathcal{P}(A_\mu \mid K_F) \geq p$$

Suppose in Table 1, P_σ needs to check for (1,p)-sufficiency. Optimal 2-anonymization for P_σ's private table T_σ is T_σ^* with $\mu(T_\sigma^*) = [0,2]$. There is only one mapping [0,1] which is 1 away from [0,2]. So we need to check if $\mathcal{P}(\Delta_{0,1}(T_\cup)$ is 2-anonymous $\mid K_F) \geq p$. Note that we do not need to check also for the mapping [0,0] since if $\Delta_{0,1}(T_\cup)$ violates k-anonymity so does $\Delta_{0,0}(T_\cup)$.

In the next section, we show how to calculate $\mathcal{P}(A_\mu \mid K_F)$, the *$\mu$-probability*, for a distributed k-anonymity protocol.

5 μ-Probability of a Protocol

Definition 12 (Bucket Set) A bucket set for a set of attributes C, and a mapping μ, is given by $B = \{\texttt{tuple}\quad b \mid \exists t$ $\texttt{from the domain of}\quad C\ \texttt{such that}\quad b^* = \Delta_\mu(t)\}$

In Table 1, for the domain tables defined and the mapping [0,1], the bucket set is given by $\{$<M,AM> ,<M,EU> ,<F,AM> , <F,EU> $\}$. When we refer to this bucket set, we will index the elements: $\{b_1, b_2, b_3, b_4\}$

5.1 Assumptions

Deriving the exact μ-probability is a computationally costly operation. To overcome this challenge, we make the following assumptions in our probabilistic model:

Attribute Independence: Until Section 5.6, we assume that there is no correlation between attributes. This is a valid assumption if we only know K_F about the unknown data. So from P_σ's point of view, for any foreign tuple $t \in T_1$; $\mathcal{P}(t[i] = v_k) = \mathcal{P}(t[i] = v_k \mid t[j] = v_\ell)$ for all $i \neq j$, v_k, and v_ℓ.

In section 5.6, we introduce bayesian networks (K_B) as a statistical information on $\bigcup_i T_i$ to capture correlations.

Tuple Independence: We assume foreign tuples are drawn from the same distribution but they are independent. Meaning for any two tuples $t_1, t_2 \in T_2$, $\mathcal{P}(t_1[i] = v_j) = \mathcal{P}(t_1[i] = v_j \mid t_2[i] = v_k)$ for all possible i, v_i, and v_k. Such equality does not necessarily hold given K_F, but for large enough data, independence is a reasonable assumption. In Section 7, we experimentally show that tuple independence assumption does not introduce any deviation from the exact μ-probability.

5.2 Deriving μ-Probability

Generalization of any table $T_\cup$ with a fixed mapping μ can only contain tuples drawn from the associated bucket set $B = \{b_1, \cdots, b_n\}$. Since we don't know $T_\cup$, the cardinality of the buckets act as a random variable. However, P_σ can

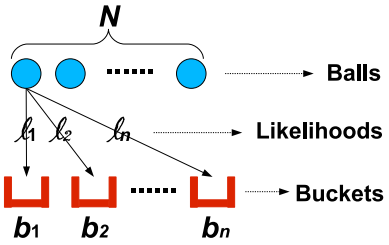


Fig. 2 Probabilistic model for μ-probability.

extract the size of the $\bigcup_i T_i$ from K_F. Letting X_i be the random variable for the cardinality of b_i, and assuming $\bigcup_i T_i$ has cardinality N, we have the constraint $\sum_i X_i = N$.

In Table 1, from P_σ's point of view $N = |T_1| = 4$. So for the four buckets above; $X_1 + X_2 + X_3 + X_4 = 4$.

The generalization $T^*_\cup$ satisfies k-anonymity if each bucket (generalized tuple) in $T^*_\cup$ has cardinality of either 0 or at least k. For horizontally partitioned data, party P_σ already knows his share on any bucket, so the buckets are initially non-empty. Let $X_i \geq^0 k$ denote the case when $(X_i = 0) \vee (X_i \geq k)$ in the case of vertically partitioned data and $X_i + |b_i \in \Delta_\mu(T_\sigma)| \geq^0 k$ in the case of horizontally partitioned data then μ-probability takes the following form:

$$\mathcal{P}(\bigcap_i X_i \geq^0 k \mid \sum_i X_i = N, K_F)$$

If we have the knowledge of the distribution functions for the attributes $K_F = \bigcup_c f_c$,the probability that a random tuple $t \in T_\cup$ will be generalized to a bucket b_i is given by [1]

$$\ell_i = \prod_c \frac{f_c(b_i[c])}{N} \tag{1}$$

which we will name as the *likelihood* of bucket b_i.

For example, in Table 1, P_σ is assumed to know the attribute distribution set $K_F = \{f^{T_2}_{\texttt{sex}}, f^{T_2}_{\texttt{nation}}\}$. (E.g., $f^{T_2}_{\texttt{sex}}$ (M) =2, $f^{T_2}_{\texttt{nation}}$ (Brazil) =1, $\cdots$). Thus the likelihood of bucket b_1 ({<M,AM>}) is $\ell_1 = \frac{f^{T_2}_{\texttt{sex}}\,\texttt{(M)}}{N} \cdot \frac{f^{T_2}_{\texttt{nation}}\,\texttt{(AM)}}{N} = \frac{2}{4} \cdot \frac{3}{4} = \frac{3}{8}$. Similarly $\ell_2 = \frac{1}{8}, \ell_3 = \frac{3}{8}, \ell_4 = \frac{1}{8}$.

Without tuple independence assumption, each X_i behaves like a hypergeometric[2] random variable with parameters (N, $N\ell_i$,N). However, hypergeometric density function is slow to compute. But with tuple independence, we can model X_i as a binomial random variable $\mathcal{B}$ [3] with parameters (N, ℓ). Such an assumption is reasonable for big N and moderate ℓ values [14]. Figure 2 summarizes our probabilistic model. Each tuple is represented by a ball with a probability ℓ_i of going into a bucket b_i. Then the μ-probability can be written as:

$$\mathcal{P}_\mu = \mathcal{P}(\bigcap_i X_i \geq^0 k \mid \sum_i X_i = N, X_i \sim \mathcal{B}(N, \ell_i)) \tag{2}$$

In Table 1, $|b_1 \in \Delta_\mu(T_\sigma)| = 2$ similarly for b_2, b_3, b_4, initial bucket sizes are $0, 1, 1$. So for $k = 2$, $\mathcal{P}_\mu = \mathcal{P}(X_1 \geq 0, X_2 \geq^0 2, X_3 \geq 1, X_4 \geq 1)$

5.3 Calculating exact μ-Probability

$\mathcal{P}_\mu$ can be calculated in two ways:

1. A recursive approach can be followed by conditioning on the last bucket:

$$\begin{aligned}
\mathcal{P}^{n,\ell_{1\cdots n}}_\mu &= \mathcal{P}(\bigcap_i^n (X_i \geq^0 k) \mid \sum_i^n X_i = N, X_i \sim \mathcal{B}(N, \ell_i)) \\
&= \sum_{x \geq^0 k} \mathcal{P}(X_n = x) \mathcal{P}(\bigcap_i^{n-1} (X_i \geq^0 k) \mid \\
&\qquad \sum_i^n X_i = N, X_i \sim \mathcal{B}(N, \ell_i), X_n = x) \\
&= \sum_{x \geq^0 k} \mathcal{B}(x; N, \ell_n) \cdot \mathcal{P}(\bigcap_i^{n-1} (X_i \geq^0 k) \mid \\
&\qquad \sum_i^{n-1} X_i = N - x, X_i \sim \mathcal{B}(N, \ell'_i)) \\
&= \sum_{x \geq^0 k} \binom{N}{x} \ell_n^x (1 - \ell_n)^{N-x} \cdot \mathcal{P}^{n-1,\ell'_{1\cdots n-1}}_\mu
\end{aligned} \tag{3}$$

where ℓ'_i is the normalized likelihood $\ell'_i = \frac{\ell_i}{\sum_j^{n-1} \ell_j}$.

2. Each tuple in $\bigcup_i T_i$ can be thought of an independent trial in a binomial process in which each trial results in exactly one of the n possible outcomes (e.g., $b_1, \cdots, b_n$). In this case, the joint random variable $(X_1, \cdots, X_n)$ follows a *multinomial distribution* with the following density function:

$$\mathcal{P}(X_1 = x_1 \cdots X_n = x_n) = \frac{N!}{x_1! \cdots x_n!} \ell_1^{x_1} \cdots \ell_n^{x_n}$$

$\mathcal{P}_\mu$ can be calculated by summing up the probabilities of all assignments that respect k-anonymity:

$$\mathcal{P}_\mu = \sum_{\sum x_i = N \wedge x_i \geq^0 k} \frac{N!}{x_1! \cdots x_n!} \ell_1^{x_1} \cdots \ell_n^{x_n} \tag{4}$$

In Table 1, following the example above, one assignment that satisfies 2-anonymity is $X_1 = 0, X_2 = 1, X_3 = 0, X_4 =$

[1] assuming attribute independence

[2] hyp(x;N,M,n):A sample of n balls is drawn from an urn containing M white and $N - M$ black balls *without replacement*. hyp gives the probability of selecting exactly x white balls.

[3] $\mathcal{B}$(x;n,p):A sample of n balls is drawn from an urn of size N containing Np white and $N(1-p)$ black balls *with replacement*. $\mathcal{B}$ gives the probability of selecting exactly x white balls.

3. The probability share of this assignment on $\mathcal{P}_\mu$ can be calculated as

$$\begin{aligned}&\mathcal{P}(X_1=0,X_2=1,X_3=0,X_4=3)\\&=\mathcal{P}(X_4=3)\mathcal{P}(X_3=0\mid X_4=3)\\&\quad\cdot\mathcal{P}(X_2=1\mid X_3=0,X_4=3)\\&\quad\cdot\mathcal{P}(X_1=0\mid X_2=1,X_3=0,X_4=3)\\&=\frac{4!}{0!\cdot1!\cdot0!\cdot3!}\ell_1^0\ell_2^1\ell_3^0\ell_4^3=0.026\end{aligned}$$

If we sum up all the probabilities for valid assignments, we get the μ probability as 0.23.

Unfortunately, calculating μ-probability with Equation 3 or 4 is computationally expensive. Number of assignments that satisfy $X_1+\cdots+X_n=N$, thus number of binomials that needs to be calculated is in the order of $\binom{N+n-1}{N}$. We next show how to bound and approximate μ-probability.

5.4 Bounding μ-Probability

Let E_i be the event that $0\leq X_i\leq k$. Obviously,

$$\mathcal{P}_\mu=\mathcal{P}(\bigcap\bar{E}_i)=1-\mathcal{P}(\bigcup E_i)$$

Bounding the probability of union of events is well studied in the literature. One of the most common bounds is given by Bonferroni [5]:

$$\mathcal{P}_\mu=1-(z_1-z_2+z_3-z_4+\cdots(-1)^{n-1}z_n)$$

where $z_1=\sum_i\mathcal{P}(E_i)$, $z_2=\sum_{i<j}\mathcal{P}(E_i,E_j)$,
$z_3=\sum_{i<j<\ell}\mathcal{P}(E_i,E_j,E_\ell),\cdots$

Note that each z_i can be calculated by applying Equation 3 on the associated X variables. Then Bonferroni bounds apply;

$$\begin{aligned}&\mathcal{P}_\mu\leq1-z_1+z_2,\ \ \mathcal{P}_\mu\leq1-z_1+z_2-z_3+z_4,\ \ \cdots\\&\mathcal{P}_\mu\geq1-z_1,\ \ \mathcal{P}_\mu\geq1-z_1+z_2-z_3,\ \ \cdots\end{aligned}$$

Following the example above, $z_1=\mathcal{P}(E_1)+\mathcal{P}(E_2)+\mathcal{P}(E_3)+\mathcal{P}(E_4)=0+\binom{4}{1}\ell_2(1-\ell_2)^3+\binom{4}{0}(1-\ell_3)^4+\binom{4}{0}(1-\ell_4)^4=1.074$. z_2=$\mathcal{P}(E_1,E_2)$+$\mathcal{P}(E_1,E_3)$+$\mathcal{P}(E_1,E_4)$+$\mathcal{P}(E_2,E_3)$ +$\mathcal{P}(E_2,E_4)$+$\mathcal{P}(E_3,E_4)$=$0+0+0+\binom{4}{1}\binom{4}{0}\ell_2(1-\ell_2-\ell_3)^3+\binom{4}{1}\binom{4}{0}\ell_4(1-\ell_2-\ell_4)^3+\binom{4}{0}\binom{4}{0}(1-\ell_3-\ell_4)^4=0.336$. So $-0,073\leq\mathcal{P}_\mu\leq0.263$.

Even though Bonferroni always holds, it does not guarantee tight bounds [25]. Besides that, calculation of high dimensional marginal distributions may still be infeasible for large data. In Section 7, we experimentally show the efficiency of the bounding algorithms.

5.5 Approximating μ-Probability

In this section, we adapt the approximation of multinomial cumulative distribution given in [17] to μ-probability. The resulting approximation is much faster to compute compared to bounding techniques. Even though the error of the approximation is unbounded, as we show in Section 7, the approximation is practically quite accurate.

Let A_i be the event that $X_i\geq^0 k$, then given $X_i\sim\mathcal{B}(N,\ell_i)$ we have;

$$\begin{aligned}\mathcal{P}_\mu&=\mathcal{P}(\bigcap A_i\mid\sum X_i=N)\\&=\frac{\mathcal{P}(\sum X_i=N\mid\bigcap A_i)\cdot\mathcal{P}(\bigcap A_i)}{\mathcal{P}(\sum X_i=N)}\\&=\frac{\mathcal{P}(\sum Y_i=N)\cdot\mathcal{P}(\bigcap A_i)}{\mathcal{P}(\sum X_i=N)}\end{aligned}\tag{5}$$

where Y_i is a truncated binomial; $Y_i\ \sim\ (X_i|X_i\geq^0 k)$.

The second numerator term is a probability of independent binomials so it can easily be computed as:

$$\mathcal{P}(\bigcap A_i)=\prod\mathcal{P}(X_i\geq^0 k)$$

The first numerator term and the denominator however is the probability regarding sums of random variables which are independent but not identically distributed. However, since both X_i and Y_i are bounded, by Lindeberg theorem [5], the central limit theorem holds; distribution of the sums converges to a normal distribution $\mathcal{N}$ as n goes to infinity. So given $(\bar{X}_i,\bar{Y}_i)$ is the mean and $(\sigma^2_{\bar{X}_i},\sigma^2_{\bar{Y}_i})$ is the variance of (X_i,Y_i) respectively, then

$$\mathcal{P}_\mu\simeq\frac{\mathcal{P}(|\mathcal{N}_Y-N|\leq0.5)}{\mathcal{P}(|\mathcal{N}_X-N|\leq0.5)}\cdot\prod\mathcal{P}(X_i\geq^0 k)$$

where $\mathcal{N}_X\sim\mathcal{N}(\sum\bar{X}_i,\sum\sigma_{X_i})$ and $\mathcal{N}_Y\sim\mathcal{N}(\sum\bar{Y}_i,\sum\sigma_{Y_i})$

Following the example in Table 1, $\sum\bar{X}_i=4$,$\sum\sigma_{X_i}=2.75$, $\sum\bar{Y}_i=4.72$,$\sum\sigma_{Y_i}=2.23$, thus approximation gives $\mathcal{P}_\mu\simeq0.23$. Approximation in this case is successful up to the third decimal even though n is considerably small.

Even though approximation is not guaranteed to lie within the Bonferroni bounds, we show in Section 7 that the approximation is very accurate in practice and very fast compared to the computation of exact algorithm and the Bonferroni bounds.

5.6 Handling Correlations

So far, we assumed that only the knowledge of the global attribute distributions is used to estimate μ-probability. However, such information cannot successfully describe a global dataset with high attribute correlations. Sharing joint attribute distribution instead of single attribute distributions can be a

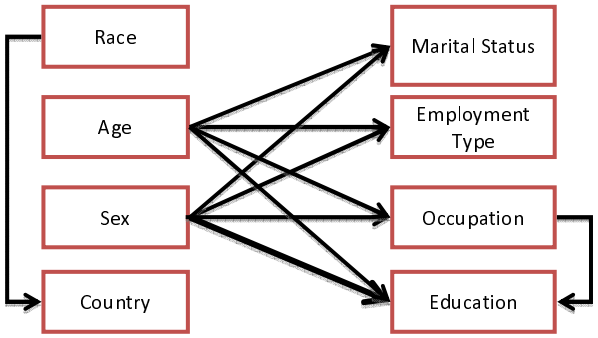


Fig. 3 A bayesian network

solution. However, such sharing should be done carefully since supplying all joint probabilities with an SMC not only is inefficient due to the large domain of joint attributes but also might give out too much sensitive information.

Instead, parties can agree on summary structures that contain information on correlations. Bayesian networks (BN) are one such example. BNs are basically directed graphs in which each vertex corresponds to an attribute and each edge shows a dependency relation between the connected attributes (see Figure 3). Attribute pairs not connected by an edge are assumed to be conditionally independent. Joint probabilities on the connected attributes are supplied with the BN structures. The advantages of using BNs are 1. the level of dependency to be shared can be adjusted, thus information disclosure and the amount of communication traffic can be limited, 2. BNs are also proposed for query selectivity in database systems, thus might be readily available within the database management system [6].

In situations where local data of a party is assumed to follow a similar distribution with the global data, BNs can be constructed locally without requiring an SMC. Such an approach is however, only applicable to horizontally partitioned data.

Incorporating information from a BN structure does not complicate the computation of μ-probability. BNs only affect the likelihoods of buckets given in Section 5.2. Thus only Equation 1 is rewritten in terms of joint distributions.

5.7 Communication Protocol

As we mentioned before, assuming we have n remote parties $P_1,\cdots,P_n$, the knowledge K_F that P_σ gets should describe $T_\cup - T_\sigma = \bigcup_{i=1}^{n} T_i$. Having P_σ get separate distributions on T_i from each party P_i would disclose too much information. Our goal in this section is two folds:

- We limit the information disclosure by extracting information from the local anonymizations rather than private tables. This is useful if the privacy policies of the participating parties prevent disclosure of any information of finer granularity.
- We use SMC protocols to calculate the global K_F. Thus for more than 2 parties, private shares inherent in the global K_F are indistinguishable. Fortunately, such a protocol is not costly for non-colluding parties.

Algorithm 1 Secure distribution of K_F

Require: Parties $\{P_\sigma, P_1,\cdots,P_n\}$, P_σ gets distribution functions on $\bigcup_{i=1}^{n} T_i$.
1: **for all** dimensions d **do**
2: Let set of values $\{v_1,\cdots v_n\}$ be the domain of d
3: P_σ sends vector of random number $R^0 = \{r_1^0,\cdots,r_n^0\}$ to P_1
4: $i = 1$
5: **while** $i < n$ **do**
6: P_i calculates distorted distribution function $G_i = \texttt{getUniform}\ (d)$
7: P_i sends $R^i = R^{i-1} + \{G_i(v_1),\cdots,G_i(v_n)\}$ to P_{i+1}
8: $i{+}{+}$
9: P_n sends R^n to P_σ
10: P_σ calculates the distribution for d; $f_d(v_i) = r_i^n - r_i^\sigma$

Algorithm 2 getUniform(d)

Require: Party has the private table T and d is a dimension in T. Set of values $V = \{v_1,\cdots v_n\}$ is the domain of dimension d.
Ensure: Function G is the distorted distribution function for dimension d.
1: Let T^* be the k-anonymization of T with mapping μ.
2: **for all** $v_i \in V$ **do**
3: Let generalized value $v^* = \Delta_\mu(v_i)$ in T^* and let p be the frequency of v^*
4: Let $V' = \{v_j \mid v_j \in V \ \wedge\ v_j \in \Delta_\mu^{-1}(v^*)\}$ and let q be the size of V'
5: Assuming each value is equally likely to appear, pick randomly a vector $R = \{r_1,\cdots,r_q\}$ such that $\sum_{i\in[1-q]} r_i = p$
6: $G(v_j) = r_j$ for $v_j \in V'$
7: $V = V - V'$
8: Return G;

Algorithm 1 shows how parties can calculate global K_F securely. In line 3, P_σ supplies a random number for each domain value v_i. In line 7, each party adds its private share (which we explain shortly) to the random sum and the last party sends the final sum back to P_σ. P_σ finds the global distribution by subtracting the initial random number from the final sum.

The important point here is that private shares of parties do not contain the exact frequency of v_i. Parties distort the frequency as given in Algorithm 2. The algorithm getUniform returns new distributions from the local k-anonymization other than the private table. The anonymized distribution (of values of possibly coarser granularity) is first extracted and a new distribution on atomic values (e.g., G) that respects the anonymized distribution is returned randomly. Randomization should enforce symmetry thus indistinguishability between each atomic value in the same

equivalence class ($\mathcal{P}(G(v_i)=x)=\mathcal{P}(G(v_j)=x)$ for all i,j,x). An example is to draw the frequencies from a multinomial distribution:

$$\mathcal{P}(r_1=x_1,\cdots,r_q=x_q \mid r_1+\cdots+r_q=p) = \begin{cases} 0, & x_1+\cdots+x_q\neq p; \\ \frac{p!}{x_1!\cdots x_q!}\frac{1}{q^p}, & \text{otherwise.} \end{cases}$$

Note that K_F from the remote parties is truthful only on a coarse granularity decided by their local anonymizations. Thus, information released by the remote parties is bounded by their local k-anonymizations.

Suppose this time we have the parties P_σ, P_1, and P_2 with private tables T_σ, T_1, and T_2 in Tables 1 and 2. We again assume P_σ initiates a look ahead with info K_F. To get the frequency of a domain value, say `Male` ; first, P_σ picks a random number r and sends it to P_1. P_1 calculates $G(\texttt{M})$ from T_1^*. Since `Sex` column is not generalized in T_1^*, $G_2(\texttt{M})$ returns the exact frequency of the males: 2. Thus, P_2 sends $r+2$ to P_2. P_2 calculates $G_2(\texttt{M})$ from T_2^*. `*` is the generalization of `M` in T_2^* with frequency 4 (e.g., $p=4$). There are two atomic values `M` and `F` under `*` (e.g., $q=2$). Thus the output of G_2 on `M` and `F` will respect the following multinomial distribution:

$$\mathcal{P}(G_2(\texttt{M})=x, G_2(\texttt{F})=y) = \begin{cases} \frac{1}{16}, x=0,y=4 \\ \frac{4}{16}, x=1,y=3 \\ \frac{6}{16}, x=2,y=2 \\ \frac{4}{16}, x=3,y=1 \\ \frac{1}{16}, x=4,y=0 \end{cases}$$

5.8 Information Disclosure

In this section, we discuss for a given party how much information on his/her private input is disclosed to other parties due to the look ahead. The amount of information disclosure depends on the outcome of the look ahead. So we evaluate the disclosure case by case. We start with disclosure by the remote parties:

*1. Insufficient SMC*If the look ahead concludes that SMC will not meet the expectations, we assume each party releases their local anonymizations. Note that Algorithm 1 only operates on the local anonymizations of the remote parties meaning any adversary can simulate the look ahead from the released anonymizations. Thus, with respect to remote parties, there is no information disclosure due to the look ahead process.

*2. Sufficient SMC*In this case parties initiate the SMC protocol. We again have two cases to consider depending on the output of the SMC protocol. Let $T_\cup^*$ be the output and party P_i has private input T_i with local anonymization T_i^*.

*2.1. $\mu(T_i^*)\subset\mu(T_\cup^*)$*If T_i^* is a higher level generalization than $T_\cup^*$; the released data $T_\cup^*$ contains more information on T_i than the local anonymization T_i^*. Thus K_F on T_i^* does not give out any more than $T_\cup^*$.

*2.2. $\mu(T_i^*)\not\subset\mu(T_\cup^*)$*We now try to upper bound the information disclosure in this case. Surely, P_i sending the exact local anonymization T_i^* to P_σ (as opposed to distributions on T_i^*) results in a higher information disclosure. In such a case, P_σ sees two different anonymizations of T_i enabling him/her conduct intersection attacks to recover some data cells in finer granularity. Table 3 shows an example where T_2^* is the local anonymization of T_2 and $T_{\cup,2}^*$ is what P_σ sees at the end of the protocol. Seeing T_2^* and $T_{\cup,2}^*$ together, P_σ can conclude that there is a tuple in T_2 with `sex=M` , `nation=EU` , `salary=>40K` . Note that such information cannot be extracted from $T_{\cup,2}^*$ alone.

However, even though it is possible, it is quite unlikely to launch intersection attacks in our protocol. Because P_σ only sees the global K_F on all local anonymizations. It is not possible to distinguish distribution of one party from that of an another. Even for the two party case, K_F is distorted and there is no way of telling the granularity of truth in K_F (e.g., generalization mapping). Besides even if P_σ knows the mapping, it is still unlikely to link values to an individual in T_i. Consider in Table 3, P_σ knows $T_{\cup,1}^*$ and distributions on T_1^*. In addition to $T_{\cup,1}^*$, P_σ will discover from the distributions that there are two nationalities of `EU` and `AM` in T_1. However, he/she will not be able to link the attribute `nation` with `sex` or `age` . In other words, what P_σ gets from K_F does not help to gain knowledge on any of the individuals.

Nevertheless, it is possible for parties to avoid case 2.2 by enforcing the output generalization mapping to be some descendant of all local mappings. This would negate any information disclosure at the cost of utility. Such an approach makes sense especially for the two party case in which disclosure risk is the highest. Also the utility loss for two party case is minimized since it becomes easier to find a common descendant mapping.

As for the disclosure by the home party, at the end of the look ahead, the remote parties will know the decision of P_σ on inputs T_σ, K_F. This information cannot be simulated from T_σ^* alone, thus there is a non-zero information disclosure. This situation could have been avoided by enforcing the home party to use T_σ^* instead of T_σ in the look ahead. However, we advocate that the risk of disclosure is very small to take such a precaution at the cost of utility. It is unlikely for the remote parties to infer anything from the decision since no party knows the exact global K_F. Besides, decision does not disclose the exact μ-probability making any inference on T_σ difficult if not impossible.

Table 2 Tables for party P_2

Name	Sex	Nation	Salary
q9	F	England	>40K
q10	M	Canada	≤40K
q11	M	USA	≤40K
q12	M	Italy	≤40K

T_2

Name	Sex	Nation	Salary
q9	*	EU	>40K
q10	*	AM	≤40K
q11	*	AM	≤40K
q12	*	EU	≤40K

T_2^*

Name	Sex	Nation	Salary
q9	F	EU	>40K
q10	M	AM	≤40K
q11	M	AM	≤40K
q12	M	EU	≤40K

$T^*_{\cup,2}$

Table 3 Intersection Attack

Name	Age	Sex	Nation	Salary
q13	12	M	Italy	>40K
q14	17	F	USA	≤40K
q15	24	M	Canada	≤40K
q16	25	F	England	≤40K

T_1

Age	Sex	Nation	Salary
10-20	*	EU	>40K
10-20	*	AM	≤40K
20-30	*	AM	≤40K
20-30	*	EU	≤40K

T_1^*

Age	Sex	Nation	Salary
10-20	M	*	>40K
10-20	F	*	≤40K
20-30	M	*	≤40K
20-30	F	*	≤40K

$T^*_{\cup,1} = T^*_{\cup} - T^*_{\sigma}$

6 Look Ahead for Distributed Recursive Diversity

6.1 Problem Definition

In this section, we show how to modify our methodology to work with distributed ℓ-diversity assuming we have different distribution functions describing tuples with different sensitive values. However, to the best of our knowledge, there is no proposed protocol for distributed ℓ-diversity problem. Thus, we choose to leave the practical evaluation of the theory as a future work.

While k-anonymity constraints on the size of the equivalence classes, recursive diversity constraints on the distribution of sensitive attributes. In this section, we try to propose an extension for recursive diversity when we have only two sensitive values. We first revise our problem definition for distributed diversity.

Definition 13 Let A_μ be the event that $\Delta_\mu(T_\cup)$ is recursive $(c,2)$ diverse. Assuming we have only two values s^1, s^2 in the sensitive attribute domain, the problem of sufficiency in this case is to prove that, for at least one $\mu \in S_\mu$;

$$\mathcal{P}(A_\mu \mid K_{F^1,F^2}) \geq p$$

where F^i is the distribution function for the set tuples with sensitive value s^i.

Note that we assume in this case, the home party collects separate distributions for each class of tuples. However by doing so, we assume independence between QI attributes with the sensitive attribute. A better approach would be to consider correlations through summary structures such as bayesian networks. We leave this challenge as a future work.

6.2 Deriving μ Probability for Recursive Diversity

The probabilistic model we construct in this section is similar to the one presented in Section 5.2. The difference is that there are now two separate distributions (one for each sensitive value s^i) of random tuples to buckets. Let N^i be the number of s^i tuples in $T_\cup$ then we have two sets of likelihood probabilities $(\bigcup_i^{N^1} \ell_i^1, \bigcup_i^{N^2} \ell_i^2)$ and cardinality random variables $(\bigcup_i^{N^1} X_i^1, \bigcup_i^{N^2} X_i^2)$, each associated with one sensitive value. μ probability for recursive (c,ℓ) diversity can be written as

Let Z_i be the event that $max(X_i^1, X_i^2) \leq c \cdot min(X_i^1, X_i^2)$, then

$$\mathcal{P}_\mu = \mathcal{P}(\bigcap Z_i \mid \overbrace{\sum_i^n X_i^1 = N^1}^{C^1}, \overbrace{X_i^1 \sim \mathcal{B}(N^1, \ell_i^1)}^{B^1},$$
$$\underbrace{\sum_i^n X_i^2 = N^2}_{C^2}, \underbrace{X_i^2 \sim \mathcal{B}(N^2, \ell_i^2)}_{B^2})$$
$$= \frac{\mathcal{P}(C^1, C^2 \mid \bigcap Z_i, B^1, B^2)\mathcal{P}(\bigcap Z_i \mid B^1, B^2)}{\mathcal{P}(C^1, C^2 \mid B^1, B^2)}$$

Since we assume QI attributes and sensitive attribute are independent, C^1, C^2 becomes independent variables. Thus,

$$\mathcal{P}_\mu = \frac{\mathcal{P}(C^1 \mid \bigcap Z_i, B^1) \cdot \mathcal{P}(C^2 \mid \bigcap Z_i, B^2)}{\mathcal{P}(C^1 \mid B^1) \cdot \mathcal{P}(C^2 \mid B^2)} \cdot \mathcal{P}(\bigcap Z_i \mid B^1, B^2)$$

$\mathcal{P}_\mu$ can be calculated or bounded with the same techniques given in Section 5.3 and 5.4. $\mathcal{P}_\mu$ can be approximated with the following equation:

$$\mathcal{P}_\mu \simeq \frac{\mathcal{P}(|\mathcal{N}_{Y^1} - N^1| \leq 0.5)}{\mathcal{P}(|\mathcal{N}_{X^1} - N^1| \leq 0.5)} \cdot \frac{\mathcal{P}(|\mathcal{N}_{Y^2} - N^2| \leq 0.5)}{\mathcal{P}(|\mathcal{N}_{X^2} - N^2| \leq 0.5)} \cdot \prod \mathcal{P}(Z_i)$$

where $\mathcal{N}_{X^j} \sim \mathcal{N}(\sum_i \bar{X}_i^j, \sum_i \sigma_{X_i^j})$ and $\mathcal{N}_Y^j \sim \mathcal{N}(\sum_i \bar{Y}_i^j, \sum_i \sigma_{Y_i^j})$ with $Y_i^j \sim X_i^j \mid Z_i$ for $j \in [1-2]$.

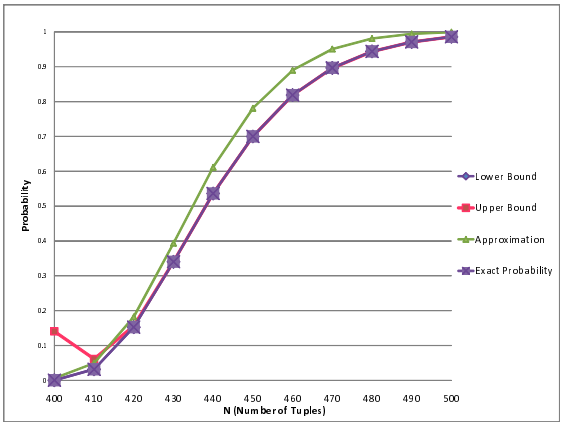


Fig. 4 Probability Results on Synthetic Dataset

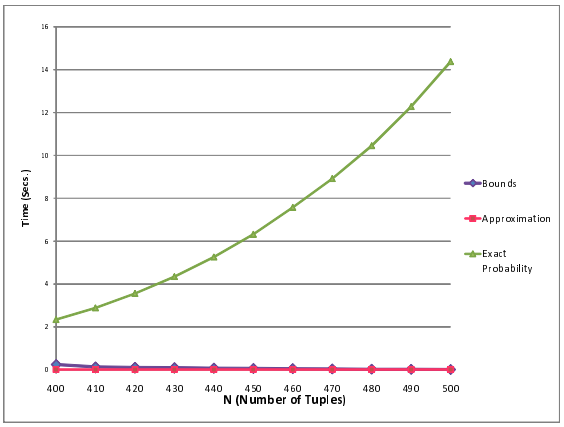


Fig. 5 Time Performance on Synthetic Dataset

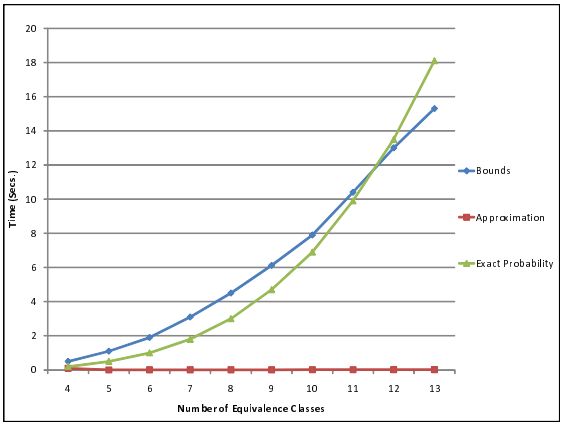


Fig. 6 Time Performance on Synthetic Dataset

7 Experiments

We evaluate our approaches in three different contexts based on the data source. In Section 7.1, we work on a synthetic uniform dataset. In this set of experiments, we vary the data size and the number of equivalence classes while uniformly distributing tuples to equivalence classes. The second and third class of experiments in Sections 7.2 and 7.3 use the famous UCI Adult dataset [2]. For the second set of experiments, we use a shuffled version of the dataset to achieve attribute independence. To be more precise, keeping the attribute distributions fixed, we shuffled the values of each attribute independently. By doing so, we set our expectation on the joint probability of a given tuple to be the product of the distributions of its values. The third set of experiments is run on the original Adult dataset with correlations between the attributes.

The experiments aim to convey the accuracy of the probability approximation and the time performance with respect to different variables such as data size and the value of k.

7.1 Synthetic Dataset

For the synthetic dataset, we set k to 100 and the number of equivalence classes to 4, which means each equivalence class will contain 25% of the total number of tuples due to the uniform distribution assumption. Figure 4 plots the results of Bonferroni Bounds (Lower and Upper) and the Approximation algorithm against the actual probability of being k-anonymous, when the size of the data varies. Bonferroni Bounds get more precise as the data size increases. Approximation tends to overestimate with a margin of at most 0.1, but seems to be independent of the data size.

The question '*Why don't we calculate the exact probability?*' is eliminated with Figure 5. The time it takes to calculate the exact probability grows exponentially with respect to the data size, whereas efficiency of Bonferroni Bounds and the Approximation seems to be independent of the data size

Although the Bonferroni Bounds seem to yield more accurate results in less time, when we increase the number of equivalence classes, the time performance of the bounds decrease drastically as shown in Figure 6. The exact probability behaves similarly whereas the time requirement of the Approximation is independent of the number of equivalence classes.

7.2 Shuffled Adult Dataset

To test our approach, we have generated an SMC scenario (Scenario 1) similar to the one in Section 2.2. We have two parties 'home site' and 'remote site' that are willing to initiate an SMC to create a global k-anonymization. The 'home site' employs the *Look Ahead* of the SMC protocol with the 'remote site', thus approximates the local info gain (Section 3). Info gain is calculated with respect to the μ-cost metric and we search for a $(1, p)$-sufficient SMC protocol. In other words, we try to look ahead to calculate the probability that

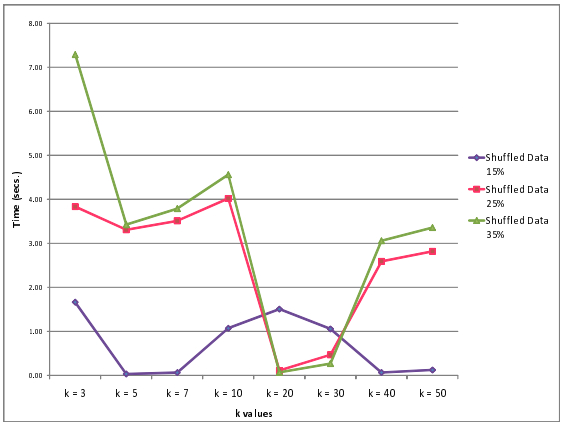


Fig. 8 Time Performance on Shuffled Dataset

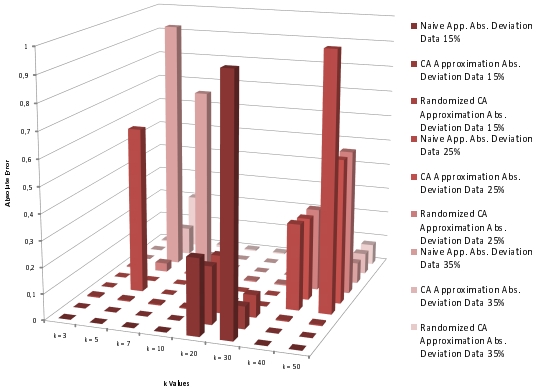


Fig. 9 Absolute Error on Adult Dataset for Scenario 1

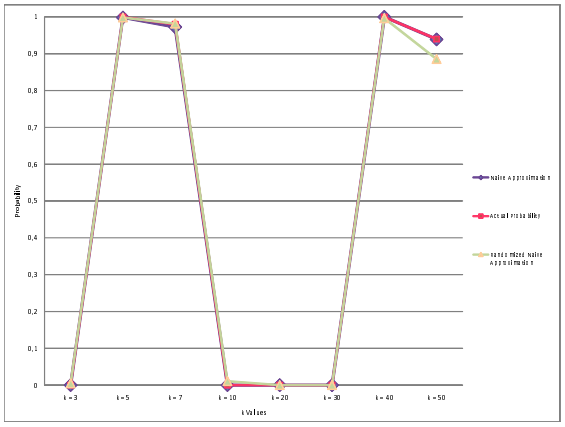
the SMC protocol returns a lower mapping generalization (a mapping positioned lower on the full domain generalization lattice) than the local generalization. As also mentioned in Corollary 1, a lower level mapping is provably more utilized than a higher level mapping. For the sake of simplicity, we check for the mapping that is one below with respect to $marital-status$ attribute.

To create the datasets for each site, we have partitioned the dataset as follows: First, we have selected and removed 15% of the data (which corresponds to 4524 tuples) to form the data of the home site. Then, using the remaining 85% of the data, by random sampling, we have formed the data of the remote site which has a size equal to the home site. Although home site knows the general distribution of the data of the remote site, data itself is invisible to the home site. Repeating the latter step for 100 times we had 100 different remote sites that are going to be subject to *Look Ahead* based on the data of the home site. Conducting several experiments on randomized data gives us an idea on the algorithm behavior at the mean. To show the effect of the data size, we have repeated the above mentioned procedure by increasing the size of the data used for both home site and remote sites to 25% of the data (7540 tuples) and 35% of the data (10556 tuples).

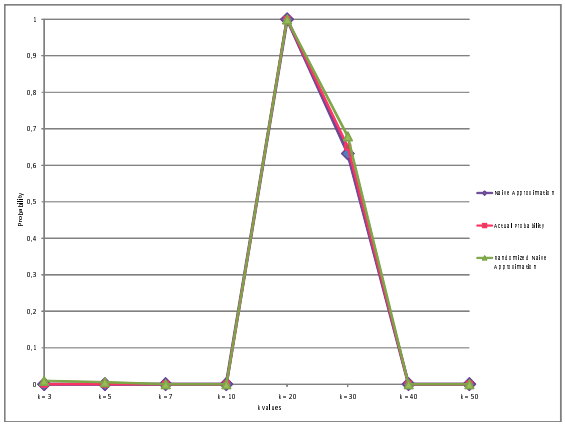
Figure 7 shows the accuracy of Naive Approximation and Randomized Naive Approximation for each data size mentioned above and for different k values. We name the approximation 'naive' as it does not take attribute correlations into account, thus Naive Approximation calculates the probability on the information gain given only attribute distributions. We use the adjective 'Randomized' when the remote site shares a distorted version of the atomic frequencies instead of the actual distribution, to bound the information released by its local k-anonymization (Section 5.7). Note that in this set of experiments, we have shuffled the data and broken the correlations between the attributes.

We have applied the algorithm against every remote site taking its corresponding home site as a base and then averaged the results to be the representative result against the actual probability. We calculate the actual probability as follows: We first combine all the local datasets to get the universal dataset (e.g., $T_{\cup}$). Then we check if k anonymity is reached with the selected mapping as described above, and take the average of the 100 runs. In other words, we try to find out the percentage of a positive information gain out of 100 cases and compare it with the probability value that the Naive Approximation returns. Experiments show that when there is a low correlation between the attributes, our results are spot on. For instance, for the 25% Data Case and for $k = 30$, our Naive Approximation algorithm finds the μ-probability as 0.65 and we found that out of 100 universal datasets, 65 of them were k-anonymous with respect to the selected mapping. This shows our probabilistic model is very successful on predicting the probability of k-anonymity when there is little correlation betweeen the attributes. Although it has a larger error margin in some cases, Randomized Naive Approximation is almost as successful as Naive Approximation. This difference obviously stems from the distortion in the information provided by the remote site. It is a trade off between accuracy and privacy but results show that we do not lose that much by disclosing less.

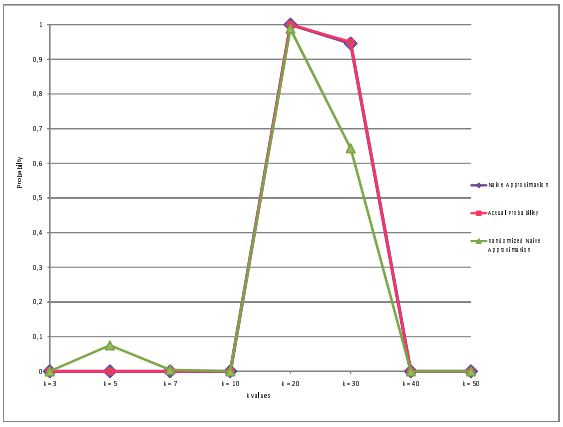
Figure 8 shows the time performance of the Naive Approximation for different k values. Time requirement for the Naive Approximation is identical to the Randomized version as the only difference between the two is the frequencies of the atomic values. Thus the time requirement of the Randomized Naive Approximation is not shown. 25% and 35% Data Cases follow similar patterns but there is no total domination between any pair of the lines. This is because the time taken by the algorithm does not depend directly on data size or k, instead it depends on the mapping used and the number of equivalence classes (e.g., buckets in Section 5) in the resulting generalization.



(a) 15% Data Case



(b) 25% Data Case



(c) 35% Data Case

Fig. 7 Accuracy of Naive and Randomized Naive Approximations on Shuffled Dataset in Different Data Sizes

7.3 Adult Dataset

In this set of experiments, we again use the Scenario 1, but this time we use the original Adult dataset, which has attribute correlations. We compare the results of the Naive Approximation and Correlation Aware Approximation (which we name as the CA Approximation from now on) that uses the Bayesian Network Structure to consider the effect of the correlations between attributes. To create data for the parties, we again sampled 15%, 25% and 35% of the Adult Dataset in the same manner as in Section 7.2, but this time we did not shuffle the data values.

Figure 9 shows the absolute error for each data case and for different k values. It can again be inferred that there is no direct relation between neither k and the absolute error nor the data size and the absolute error. We rather expect a relation between the mapping and the absolute error. We see clearly that Naive Approximation fails to capture the correlations and yields results that are off the target. For instance, in the 15% Data Case, for $k = 30$, Naive Approximation gives a μ-probability of 1, but out of 100 unions

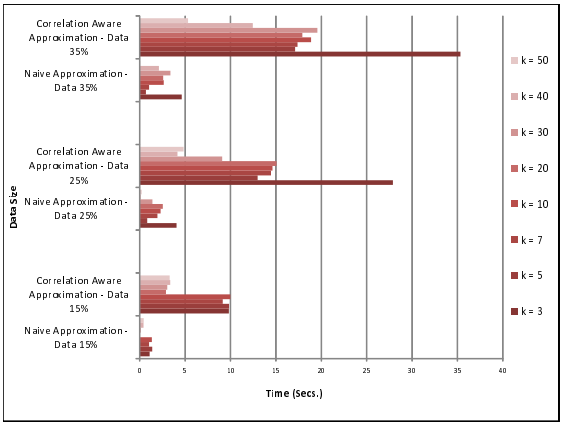


Fig. 10 Time Performance on Adult Dataset

of the home site with the remote sites, only 4 of them are k-anonymous with respect to the given mapping μ. On the other hand, the CA Approximation dominates the Naive Approximation in all cases. It yields a μ-probability of 0.12 for the above mentioned case and has an absolute error of 0.08,

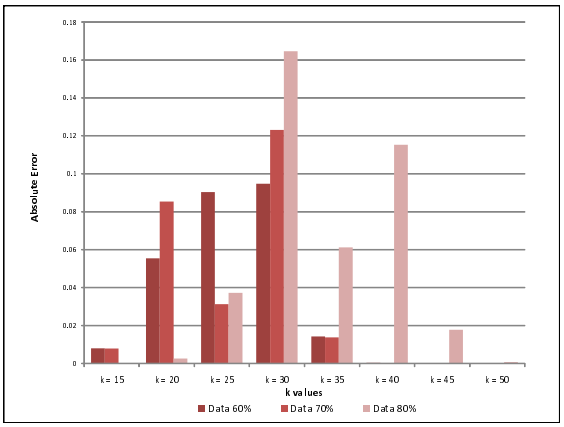


Fig. 11 Absolute Error on Adult Dataset for Scenario 2

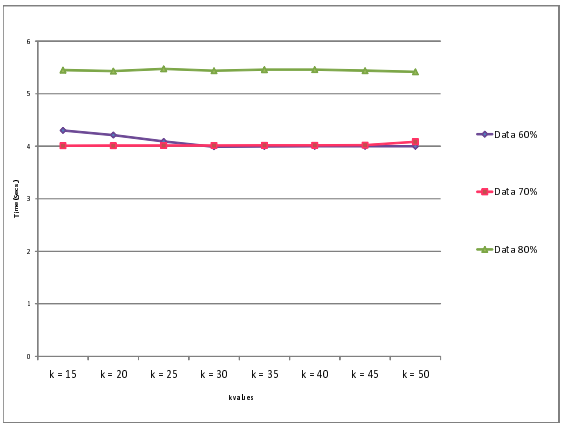


Fig. 12 Time Performance on Adult Dataset for Scenario 2

compared to the absolute error of 0.96 of the Naive Approximation. The performance of Randomized CA Approximation is very close to the CA Approximation's performance, just like in the Naive Approximation case in Section 7.2. We are off the mark in some cases as we trade privacy for accuracy, but generally the results are as accurate as CA Approximation.

Figure 10 shows the time performance for the test mentioned above. Again the time performances of Randomized CA Approximation is identical to CA Approximation and is not shown in the figure. The value of k again does not directly affect the time requirement of the algorithm, but the main decisive factor is the mapping used.

We have generated a new scenario (Scenario 2), where we do not assume any parties thus do not partition the data and we try to approximate the probability of being k anonymous looking at a single sample from the original dataset for a fixed mapping. Our aim is to work with a larger dataset and factor out the effect of the generalization mapping on accuracy and efficiency. We have sampled 100 new and independent datasets of size 60% of the original set. We have repeated the same procedure while changing the sample data size to 70% and 80% of the original dataset.

Figure 11 shows the absolute error of CA Approximation for each data case and for different k values. We see that the error rate is at most 0.17. Although both columns have similar shapes, 80% Data Case seems to be shifted to right. That is because there is no direct relation between the k value and the error rate, rather there is a relation between the number and size of equivalence classes in data and the error rate.

Finally, Figure 12 shows the time performance of CA approximation. We again see that time taken depends on the equivalence classes in the resulting generalization rather than the data size and the k value, as 60% and 70% Data Cases have similar structures and 80% Data Case requires more time. Note that the time required for a CA approximation is no more than 6 seconds. Compared to the running time of a semi honest SMC protocol where execution time is generally expressed in days or weeks, the look ahead approach is quite cheap and effective.

8 Conclusion and Future Work

Most SMC protocols are expensive in both communication and computation. We introduced a look ahead approach for SMC protocols that helps involved parties to decide whether the protocol will meet the expectations before initiating it. We presented a look ahead specifically for the distributed k-anonymity by approximating the probability that the output of the SMC will be more utilized than their local anonymizations. Experiments on real data showed that the look ahead process is perfectly accurate given non-identifying statistics on the global union.

Designing look aheads for other SMC protocols stands as a future work. A wide variety of SMC protocols have been proposed especially for privacy preserving data mining applications [19,28,12] each requiring a unique look ahead approach. As for the look ahead process on distributed anonymization protocols, definitions of k-anonymity definitions can be revisited, more efficient techniques can be developed and experimentally evaluated.

References

1. R. J. Bayardo and R. Agrawal. Data privacy through optimal k-anonymization. In *ICDE'05: Proceedings of the 21st International Conference on Data Engineering*, pages 217–228, Washington, DC, USA, 2005. IEEE Computer Society.
2. C. Blake and C. J. Merz. UCI repository of machine learning databases. `http://www.ics.uci.edu/~mlearn/MLRepository.html`, University of California, Irvine, Department of Information and Computer Sciences.
3. B.-C. Chen, K. LeFevre, and R. Ramakrishnan. Privacy skyline: privacy with multidimensional adversarial knowledge. In *VLDB*

'07: Proceedings of the 33rd international conference on Very large data bases, pages 770–781. VLDB Endowment, 2007.

4. J. Domingo-Ferrer and V. Torra. Ordinal, continuous and heterogeneous k-anonymity through microaggregation. *Data Mining and Knowledge Discovery*, 11(2):195–212, 2005.
5. W. Feller. *An Introduction to Probability Theory and Its Applications, Volume 1*. Wiley, 1968.
6. L. Getoor, B. Taskar, and D. Koller. Selectivity estimation using probabilistic models. *SIGMOD Rec.*, 30(2):461–472, 2001.
7. O. Goldreich. *The Foundations of Cryptography*, volume 2, chapter General Cryptographic Protocols. Cambridge University Press, 2004.
8. A. O. hrn and L. Ohno-Machado. Using boolean reasoning to anonymize databases. *Artificial Intelligence in Medicine*, 15(3):235–254, Mar. 1999.
9. V. S. Iyengar. Transforming data to satisfy privacy constraints. In *KDD'02: Proceedings of the eighth ACM SIGKDD International Conference on Knowledge Discovery and Data Mining*, pages 279–288, New York, NY, USA, 2002. ACM.
10. W. Jiang and C. Clifton. Privacy-preserving distributed *k*-anonymity. In *Proceedings of the 19th Annual IFIP WG 11.3 Working Conference on Database and Applications Security*, Storrs, Connecticut, Aug. 7-10 2005.
11. W. Jiang and C. Clifton. A secure distributed framework for achieving *k*-anonymity. *Special Issue of the VLDB Journal on Privacy-Preserving Data Management*, Sept. 2006.
12. M. Kantarcıoğlu and C. Clifton. Privacy-preserving distributed mining of association rules on horizontally partitioned data. *IEEE Transactions on Knowledge and Data Engineering*, 16(9):1026–1037, Sept. 2004.
13. D. Kifer and J. Gehrke. Injecting utility into anonymized datasets. In *SIGMOD'06: Proceedings of the 2006 ACM SIGMOD International Conference on Management of Data*, pages 217–228, New York, NY, USA, 2006. ACM Press.
14. S. N. Lahiri, A. Chatterjeea, and T. Maiti. Normal approximation to the hypergeometric distribution in nonstandard cases and a sub-gaussian berryesseen theorem. *Journal of Statistical Planning and Inference*, 137(11):3570–3590, Nov. 2007.
15. K. LeFevre, D. J. DeWitt, and R. Ramakrishnan. Incognito: Efficient full-domain k-anonymity. In *SIGMOD'05: Proceedings of the 2005 ACM SIGMOD International Conference on Management of Data*, pages 49–60, New York, NY, USA, 2005. ACM.
16. K. LeFevre, D. J. DeWitt, and R. Ramakrishnan. Mondrian multidimensional k-anonymity. In *ICDE'06: Proceedings of the 22nd International Conference on Data Engineering*, pages 25–35, Atlanta, GA, Apr. 3-7 2006.
17. B. Levin. A representation for multinomial cumulative distribution functions. *The Annals of Statistics*, 9(5):1123–1126, 1981.
18. N. Li and T. Li. t-closeness: Privacy beyond k-anonymity and l-diversity. In *ICDE'07: Proceedings of the 23nd International Conference on Data Engineering*, Istanbul, Turkey, Apr. 16-20 2007.
19. Y. Lindell and B. Pinkas. Privacy preserving data mining. In *Journal of Cryptology*, pages 36–54. Springer-Verlag, 2000.
20. A. Machanavajjhala, J. Gehrke, D. Kifer, and M. Venkitasubramaniam. ℓ-diversity: Privacy beyond *k*-anonymity. In *ICDE'06: Proceedings of the 22nd IEEE International Conference on Data Engineering*, Atlanta Georgia, Apr. 2006.
21. D. J. Martin, D. Kifer, A. Machanavajjhala, J. Gehrke, and J. Y. Halpern. Worst-case background knowledge for privacy-preserving data publishing. In *ICDE'07: Proceedings of the 23nd International Conference on Data Engineering*, Istanbul, Turkey, Apr. 16-20 2007.
22. M. E. Nergiz, M. Atzori, and C. Clifton. Hiding the presence of individuals in shared databases. In *SIGMOD'07: Proceedings of the 2007 ACM SIGMOD International Conference on Management of Data*, Beijing, China, June 11-14 2007.
23. M. E. Nergiz and C. Clifton. Thoughts on k-anonymization. *Data and Knowledge Engineering*, 63(3):622–645, Dec. 2007.
24. P. Samarati. Protecting respondent's privacy in microdata release. *IEEE Transactions on Knowledge and Data Engineering*, 13(6):1010–1027, Nov./Dec. 2001.
25. S. J. Schwager. Bonferroni sometimes loses. *The American Statistician*, 38(3):192–197, 1984.
26. L. Sweeney. Achieving *k*-anonymity privacy protection using generalization and suppression. *International Journal on Uncertainty, Fuzziness and Knowledge-based Systems*, 10(5), 2002.
27. L. Sweeney. k-Anonymity: A model for protecting privacy. *International Journal of Uncertainty, Fuzziness Knowledge-Based Systems*, 10(5):557–570, 2002.
28. J. Vaidya. *Privacy Preserving Data Mining over Vertically Partitioned Data*. PhD thesis, Department of Computer Sciences, Purdue University, West Lafayette, Indiana, 2004.
29. R. C.-W. Wong, J. Li, A. W.-C. Fu, and K. Wang. (α, k)-anonymity: An enhanced k-anonymity model for privacy preserving data publishing. In *KDD'06: Proceedings of the 12th ACM SIGKDD International Conference on Knowledge Discovery and Data Mining*, pages 754–759, New York, NY, USA, 2006. ACM.
30. X. Xiao and Y. Tao. M-invariance: Towards privacy preserving re-publication of dynamic datasets. In *SIGMOD'07: Proceedings of the 2007 ACM SIGMOD International Conference on Management of Data*, pages 689–700, New York, NY, USA, 2007. ACM.
31. S. Zhong, Z. Yang, and R. N. Wright. Privacy-enhancing k-anonymization of customer data. In *PODS'05: Proceedings of the 24th ACM SIGMOD-SIGACT-SIGART Symposium on Principles of Database Systems*, pages 139–147, New York, NY, USA, 2005. ACM Press.